The O-PLS methodology for orthogonal signal correction
- is it correcting or confusing?

Ulf G. Indahl[†]

January 2, 2017

[†]) *Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, N-1432 Ås, NORWAY*

Email: ulf.indahl@nmbu.no

Abstract

The separation of predictive and non-predictive (or orthogonal) information in linear regression problems is considered to be an important issue in Chemometrics. Approaches including net analyte preprocessing (NAP) methods and various orthogonal signal correction (OSC) methods have been studied in a considerable number of publications. In the present paper we focus on the simplest single response versions of some of the early OSC-approaches including Fearns OSC, the O-PLS, the target projection (TP) and the PLS post-processing by similarity transformation (PLS+ST). These methods are claimed to yield improved model building and interpretation alternatives compared to ordinary PLS, by filtering "off" the response-orthogonal parts of the samples in a dataset. We point at some fundamental misconceptions that were made in the justification of the PLS-related OSC-algorithms, and explain the key properties of the resulting modelling.

Keywords: OSC; O-PLS; TP, PLS+ST, NAP.

1 Introduction

The concept of orthogonal signal correction (OSC) with focus on applications to near-infrared (NIR) spectra was introduced by Wold et al. [1]. Its motivation is taken from the fact that the spectra representing the samples of a particular dataset often are contaminated by systematic variation that is unrelated to the measured responses. The purpose of OSC as implemented in the equivalent methods *Orthogonal projections to latent structures* (O-PLS) [2], *PLS post-processing by similarity transformation* (PLS+ST) [4] and the *Target projection* (TP) method [5], is to identify and eliminate so-called *orthogonal variation* in a dataset to achieve better models and/or interpretations in multivariate calibration.

The goal of the present paper is to discuss and make it even clearer how these OSC-methods work, and to give a rigorous explanation of why the entire OSC-concept may be both confusing and superfluous. To be able to follow the given arguments, familiarity with some simple undergraduate linear algebra is required. Concepts such as

- Orthogonality
- Vector space basis
- Matrix rank
- Projections onto subspaces
- The normal equations of ordinary least squares (OLS) regression
- The Gram-Schmidt orthogonalization process and the associated QR-factorization
- The singular value decomposition (SVD) of a matrix

are all assumed to be familiar. We also assume the mathematical equivalence between the various PLS algorithms studied in [3] to be known (i.e. that the numerical differences between models produced by these algorithms are only due to truncation errors caused by floating-point arithmetic). References to PLS modelling will therefore focus on its mathematical properties only, with the exception that we stress the importance of the $\mathbf{y}$-deflation that is often omitted.

The mathematical notation used below is mainly consistent with the standard chemometrics notation for the PLS methodology. We will restrict ourselves to the *single response* case, i.e. $\mathbf{y} \in \mathbb{R}^n$ is a column vector, and the corresponding $(n \times p)$ data matrix $\mathbf{X}$ has n rows associated with the number of samples, and p columns associated with the number of predictors.

In multiple linear regression modelling, the essentials of $\mathbf{y}$-orthogonality can be understood by inspecting Figure 1. It is well known that in a k-dimensional $(1 < k \leq n)$ subspace $V \subseteq \mathbb{R}^n$ for approximating $\mathbf{y}$ (where V is spanned by the columns of $\mathbf{X}$, or a set of linear combinations of these columns as in PLS and principal component regression (PCR)), the orthogonal projection $\hat{\mathbf{y}}$ accounts for exactly one dimension.

The remaining $(k-1)$ dimensions of V is spanned by $(k-1)$ additional vectors, and all of them can be selected to be orthogonal to $\hat{\mathbf{y}}$. Together with $\hat{\mathbf{y}}$ these additional vectors represent a basis

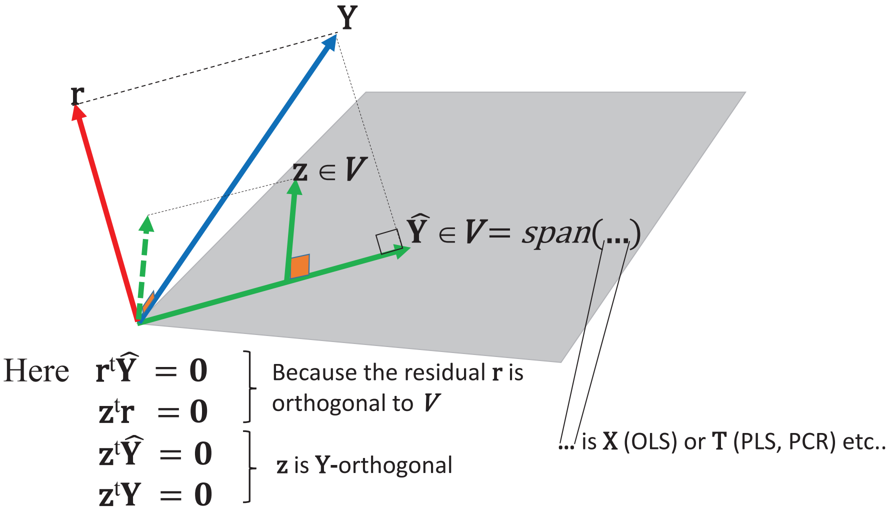
Y
r
z ∈ V
Ŷ ∈ V= span(...)
Here r^tŶ = 0
z^tr = 0
z^tŶ = 0
z^tY = 0
Because the residual r is orthogonal to V
z is Y-orthogonal
... is X (OLS) or T (PLS, PCR) etc..

Figure 1: The orthogonal projection $\hat{\mathbf{y}}$ of $\mathbf{y}(= \hat{\mathbf{y}} + \mathbf{r})$ onto the predictor subspace V occupies one dimension, and the residual vector $\mathbf{r}$ is by definition orthogonal to the subspace V.

for the subspace V. Because the residual vector $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to V, any vector $\mathbf{z} \in V$ orthogonal to $\hat{\mathbf{y}}$ is also orthogonal to $\mathbf{y} = \hat{\mathbf{y}} + \mathbf{r}$ (as a sum of two vectors both being orthogonal to $\mathbf{z}$). This observation means that a $(k-1)$-dimensional subspace of V is spanned by the $\mathbf{y}$-orthogonal vectors. Note that any such $\mathbf{y}$-orthogonal vector $\mathbf{z} \in V$ can always be represented as a linear combination of the columns in $\mathbf{X}$, i.e. $\mathbf{z} = \mathbf{X}\mathbf{w}$ for some appropriate $\mathbf{w} \in \mathbb{R}^p$.

Exhaustive descriptions of particular OSC- or partial least squares (PLS) algorithms will not be reproduced in the present paper, but a careful inspection and understanding of Figure 1 will provide the reader with a flying start to understanding the essential parts of the various OSC-methodologies discussed below.

2 The definition of y-orthogonal information in OSC

The fundamental idea in [1] was to demonstrate that an appropriate modification of the PLS algorithm can eliminate systematic $\mathbf{y}$-orthogonal parts in a data matrix $\mathbf{X}$. In [1, section 5] attention was drawn towards the possibility of identifying unit length vectors of *weights* $\mathbf{w}$ ($\|\mathbf{w}\| = 1$) were the corresponding vectors of *scores* $\mathbf{t} = \mathbf{X}\mathbf{w}$ were required to be orthogonal to $\mathbf{y}$, i.e.

$$\mathbf{t}^t\mathbf{y} = \mathbf{0}, \tag{1}$$

together with the following motivating explanation:

"*...Hence the OSC algorithm will be identical to the ordinary PLS algorithm except for the crucial step of calculating the weights* $\mathbf{w}$. *Normally, these are calculated as to maximize the covariance between* $\mathbf{X}$ *and* $\mathbf{Y}$, *but here they will instead be calculated as to minimize this covariance, i.e., to get as close to orthogonality between* $\mathbf{t}$ *and* $\mathbf{Y}$ *as possible.*"[1]

Equation (2) below is a direct extension of equation (1) showing that the $\mathbf{y}$-*orthogonal information* associated with the samples/spectra (the rows of $\mathbf{X}$) is directly related to the weights $\mathbf{w}$ defining the $\mathbf{y}$-orthogonal scores:
For any $\mathbf{y}$-orthogonal score vector $\mathbf{t} = \mathbf{X}\mathbf{w}$ we have

$$\mathbf{t}^t\mathbf{y} = (\mathbf{X}\mathbf{w})^t\mathbf{y} = \mathbf{w}^t\mathbf{X}^t\mathbf{y} = \mathbf{w}^t(c\mathbf{w}_1) = \mathbf{0} \Leftrightarrow \mathbf{w}^t\mathbf{w}_1 = \mathbf{0}, \tag{2}$$

where the unit vector $\mathbf{w}_1 = c^{-1}\mathbf{X}^t\mathbf{y}$ is the first loading weight vector of ordinary PLS, and $c = \|\mathbf{X}^t\mathbf{y}\|$ is the required normalization constant. Equation (2) says that $\mathbf{y}$-orthogonality of $\mathbf{t} = \mathbf{X}\mathbf{w}$ is equivalent to $\mathbf{w}_1$-orthogonality for the associated weight vector $\mathbf{w}$ defining the score vector $\mathbf{t}$.

It is well known that for a k-component PLS model, the associated matrix of unit loading weights $\mathbf{W} = [\mathbf{w}_1 \ \mathbf{w}_2 \ ... \ \mathbf{w}_k]$ has orthogonal orthogonal columns, i.e. $\mathbf{W}^t\mathbf{W} = \mathbf{I}$ (the identity matrix). In particular, the vectors $\mathbf{w}_2, ..., \mathbf{w}_k$ are all orthogonal to $\mathbf{w}_1$. Consequently the corresponding non-orthogonal PLS scores

$$\mathbf{t}_i^\star = \mathbf{X}\mathbf{w}_i \text{ for } 2 \leq i \leq k, \tag{3}$$

are all $\mathbf{y}$-orthogonal! This observation stands in stark contrast to the above quote from [1, section 5]. The fact that subsequent PLS components maximize the covariance between $\mathbf{X}$ and $\mathbf{y}$ only in the deflated sense of $\mathbf{y}$ (i.e. the residual $\mathbf{y}$'s) seems to have escaped the attention of both the authors and referees.

One should note that deflation of $\mathbf{X}$ with respect to the vector $\mathbf{w}_1$ yields the matrix

$$\tilde{\mathbf{X}} = \mathbf{X} - \mathbf{X}\mathbf{w}_1\mathbf{w}_1^t, \tag{4}$$

of rank one less than $\mathbf{X}$. Clearly, the rows of $\tilde{\mathbf{X}}$ are $\mathbf{w}_1$-orthogonal by construction. The $\tilde{\mathbf{X}}$-columns are all $\mathbf{y}$-orthogonal because

$$\tilde{\mathbf{X}}^t\mathbf{y} = \mathbf{X}^t\mathbf{y} - \mathbf{w}_1\mathbf{w}_1^t\mathbf{X}^t\mathbf{y} = c\mathbf{w}_1 - \mathbf{w}_1\mathbf{w}_1^t(c\mathbf{w}_1) = c\mathbf{w}_1 - c\mathbf{w}_1 = \mathbf{0}. \tag{5}$$

The deflation in equation (4) is precisely the first deflation step in the non-orthogonal scores PLS-algorithm of Martens (see [3]), and this algorithm calculates both the loading weights in $\mathbf{W}$ and the non-orthogonal scores in (3). Both Ergon [4] and Kemsley and Tapp [6] has earlier emphasized the $\mathbf{y}$-orthogonal property of the non-orthogonal PLS scores.

The various suggested algorithms and early applications [7]-[11] for doing OSC all concentrate on $\mathbf{y}$-orthogonality in the column space of $\mathbf{X}$ based on several alternative filterings of the samples. As pointed out in [10] and [12], the OSC methods are closely related to the net analyte preprocessing (NAP) approach of Goicoechea and Olivieri [11], and there is an exact algorithmic equivalence between the NAP and the direct orthogonalization (DO) method suggested by Andersson [8]. However, none of these papers consider the $\mathbf{y}$-orthogonality of the non-orthogonal PLS-scores in (3).

The NAP/DO approach derive a set of loading weights (not necessarily contained in the row space of $\mathbf{X}$) representing phenomena considered to be irrelevant to the modelling of $\mathbf{y}$ as follows:

1. Project $\mathbf{X}$ onto the orthogonal complement of the subspace spanned by $\mathbf{y}$ to obtain

$$\hat{\mathbf{X}} = \mathbf{X} - \mathbf{y}(\mathbf{y}^t\mathbf{y})^{-1}\mathbf{y}^t\mathbf{X} = (\mathbf{I} - \mathbf{y}(\mathbf{y}^t\mathbf{y})^{-1}\mathbf{y}^t)\mathbf{X}.$$

2. The columns of $\hat{\mathbf{X}}$ are clearly $\mathbf{y}$-orthogonal, i.e. $\mathbf{y}^t\hat{\mathbf{X}} = \mathbf{0}$ (but they are not necessarily contained in the column space of $\mathbf{X}$).

3. The (a) most dominant right singular (unit) vectors $\mathbf{P}_a$ from the SVD of $\hat{\mathbf{X}}$ are taken to represent the irrelevant $\mathbf{y}$-orthogonal phenomena somehow present in the data $\mathbf{X}$.

4. The NAP/DO corrected data $\mathbf{X}^\star = \mathbf{X} - \mathbf{X}\mathbf{P}_a\mathbf{P}_a^t = \mathbf{X}(\mathbf{I} - \mathbf{P}_a\mathbf{P}_a^t)$ represent a row-projection of $\mathbf{X}$ onto the orthogonal complement of the subspace spanned by $\mathbf{P}_a$, i.e the $\mathbf{X}$-data are "blinded" w.r.t. the $\mathbf{P}_a$-directions that are assumed to account for irrelevant information in the particular modelling of $\mathbf{y}$.

5. The modelling as well as subsequent model applications to new data is recommended to be based on an initial "blinding" of the datapoints by using the row projection $(\mathbf{I} - \mathbf{P}_a\mathbf{P}_a^t)$ as indicated above.

The "blinding" part in step 4. above is essential. It assures that any resulting vector of regression coefficients (say $\mathbf{b}$) obtained as a linear combination of the ($\mathbf{P}_a$-orthogonal) rows in $\mathbf{X}^\star$, is also orthogonal to the irrelevant phenomena (interferents) represented by $\mathbf{P}_a$, i.e. $\mathbf{P}_a^t\mathbf{b} = \mathbf{0}$. Note that also the following holds:

$$\mathbf{X}^\star\mathbf{b} = \mathbf{X}(\mathbf{I} - \mathbf{P}_a\mathbf{P}_a^t)\mathbf{b} = \mathbf{X}\mathbf{b} - \mathbf{X}\mathbf{P}_a(\mathbf{P}_a^t\mathbf{b}) = \mathbf{X}\mathbf{b} - \mathbf{0} = \mathbf{X}\mathbf{b}. \tag{6}$$

Equation (6) says that the application of the model represented by $\mathbf{b}$ does not require the data to be "blinded" w.r.t. the $\mathbf{P}_a$-directions, because the blinding is already taken into account in the regression coefficients $\mathbf{b}$. Preprocessing of future data points, by the "blinding" projection $(\mathbf{I} - \mathbf{P}_a\mathbf{P}_a^t)$ in applications of the model, is therefore superfluous.

Finally, we note that if $\mathbf{P}_a$ is a matrix of (a) apriori known and highly reliable interferents (not necessarily obtained from the SVD of the present $\hat{\mathbf{X}}$-matrix) for the particular modelling problem, the steps 1.-3. above could be ignored. When doing the modelling directly based on the $\mathbf{P}_a$-blinded data $\mathbf{X}^\star$ derived in step 4, the $\mathbf{P}_a$-orthogonality of the regression coefficients $\mathbf{b}$ and equation (6) still holds. Therefore, new (raw) data points can be applied "unblided" with the regression model.

2.1 Fearns OSC alternative

With reference to the definition of $\mathbf{y}$-orthogonality in [1], Fearn [13] proposed finding good $\mathbf{y}$-orthogonal pairs $(\mathbf{t}, \mathbf{w})$ by maximizing the squared norm $\|\mathbf{X}\mathbf{w}\|^2$ subject to the requirements $\|\mathbf{w}\| = 1$ and

$$\mathbf{t}^t\mathbf{y} = \mathbf{w}^t\mathbf{X}^t\mathbf{y} = \mathbf{0}. \tag{7}$$

By (2) the requirement (7) is equivalent to

$$\mathbf{w}^t\mathbf{w}_1 = 0. \tag{8}$$

This observation slightly simplifies some of the notation in [13], and shows that with $\mathbf{w}_1$ being the unit vector proportional to $\mathbf{X}^t\mathbf{y}$ (i.e. $\mathbf{w}_1$ is identical to the first PLS vector of loading weights), the maximization problem stated by Fearn corresponds to finding the dominant right singular vector of the $\mathbf{w}_1$-deflated matrix $\tilde{\mathbf{X}}$ defined in (4). Additional factors are given by the subsequent right singular vectors of $\tilde{\mathbf{X}}$ ordered by the associated singular values. The singular value decomposition (SVD), or equivalently the principal component analysis (PCA), of $\tilde{\mathbf{X}}$ therefore defines the desired maximum variance $\mathbf{y}$-orthogonal factors.

Now, assume that the desired $f(\geq 1)$ right singular vectors of $\tilde{\mathbf{X}}$ are denoted $\mathbf{w}_2,\ \mathbf{w}_3, ..., \mathbf{w}_{k(=f+1)}$ (here it is helpful to start the vector indexing from 2 and to define $k = f+1$). From the definition of $\tilde{\mathbf{X}}$ in (4) and its associated singular value decomposition, it follows that the corresponding $\mathbf{y}$-orthogonal scores $\mathbf{t}_i = \mathbf{X}\mathbf{w}_i = \tilde{\mathbf{X}}\mathbf{w}_i$ $(2 \leq i \leq k)$ are also mutually orthogonal (they are scaled versions of the left singular vectors of $\tilde{\mathbf{X}}$). The corresponding vectors of loadings are given by $\mathbf{p}_i = \mathbf{X}^t\mathbf{t}_i/(\mathbf{t}_i^t\mathbf{t}_i)$, and if we define

$$\mathbf{W} = [\mathbf{w}_2\ ...\ \mathbf{w}_k], \tag{9}$$

the associated filtered data matrix is given by

$$\mathbf{X}_o = \mathbf{X} - \mathbf{T}\mathbf{P}^t. \tag{10}$$

Here, the $\mathbf{y}$-orthogonal scores $\mathbf{T} = \mathbf{X}\mathbf{W} = [\mathbf{t}_2\ ...\ \mathbf{t}_k]$ and corresponding loadings $\mathbf{P}^t = (\mathbf{T}^t\mathbf{T})^{-1}\mathbf{T}^t\mathbf{X} = [\mathbf{p}_2\ ...\ \mathbf{p}_k]^t$. Note that with these definitions we have $\mathbf{X}_o^t\mathbf{y} = \mathbf{X}^t\mathbf{y} = c\mathbf{w}_1$ and $\mathbf{P}^t\mathbf{W} = \mathbf{I}$.

For regression purposes (see section 2.1.1), Fearns approach boils down to computing the fitted values from a final scaling and projection of $\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1$ onto the orthogonal complement of the selected $\mathbf{y}$-orthogonal left singular vectors accounting for the dominant variance in $\tilde{\mathbf{X}}$.

2.1.1 Regression modelling details of Fearns approach

For regression purposes the score vector $\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1$ (identical to the first PLS score vector) and its filtered version $\mathbf{t}_o$ with respect to the $\mathbf{y}$-orthogonal scores $\mathbf{T}$ is considered, i.e.

$$\begin{aligned}\mathbf{t}_o = \mathbf{X}_o\mathbf{w}_1 = (\mathbf{X} - \mathbf{T}\mathbf{P}^t)\mathbf{w}_1 = \mathbf{X}\mathbf{w}_1 - \mathbf{T}\mathbf{P}^t\mathbf{w}_1 = \mathbf{t}_1 - \mathbf{T}(\mathbf{T}^t\mathbf{T})^{-1}\mathbf{T}^t\mathbf{X}\mathbf{w}_1 \\ = \mathbf{t}_1 - \mathbf{T}(\mathbf{T}^t\mathbf{T})^{-1}\mathbf{T}^t\mathbf{t}_1 = (\mathbf{I} - \mathbf{T}(\mathbf{T}^t\mathbf{T})^{-1}\mathbf{T}^t)\mathbf{t}_1.\end{aligned} \tag{11}$$

Equation (11) shows that $\mathbf{t}_o$ is obtained by a Gram-Schmidt step projecting $\mathbf{t}_1$ onto the orthogonal complement of the subspace spanned by the chosen $f = k-1$ first left singular vectors of the $\mathbf{y}$-orthogonal matrix $\tilde{\mathbf{X}}$.

By defining the fitted values $\hat{\mathbf{y}}$ as the appropriately scaled version of $\mathbf{t}_o$, i.e. $\hat{\mathbf{y}} = \alpha\mathbf{t}_o = \mathbf{X}_o(\alpha\mathbf{w}_1)$ (where the scalar α is the least squares solution of $a\mathbf{t}_o = \mathbf{y}$), the corresponding $\mathbf{X}_o$-regression coefficients are

$$\boldsymbol{\beta}_o = \alpha\mathbf{w}_1. \tag{12}$$

The latter means that $\boldsymbol{\beta}_o$ is always a scaled version of the first weight vector $\mathbf{w}_1$ obtained by traditional PLS modelling.

To compute the corresponding $\mathbf{X}$-regression coefficients $\boldsymbol{\beta}$ (associated with the original unfiltered measurements), we first note that there is an alternative useful expression for $\mathbf{t}_o$, i.e.

$$\mathbf{t}_o = \mathbf{X}_o\mathbf{w}_1 = \mathbf{X}\mathbf{w}_1 - \mathbf{X}\mathbf{W}\mathbf{P}^t\mathbf{w}_1 = \mathbf{X}(\mathbf{I} - \mathbf{W}\mathbf{P}^t)\mathbf{w}_1. \tag{13}$$

Thus, we also have

$$\hat{\mathbf{y}} = \mathbf{X}_o\boldsymbol{\beta}_o = \mathbf{X}\boldsymbol{\beta}, \text{ where } \boldsymbol{\beta} = (\mathbf{I} - \mathbf{W}\mathbf{P}^t)\boldsymbol{\beta}_o. \tag{14}$$

Equation (14) shows that rather than filtering the present $\mathbf{X}$ matrix (as well as new $\mathbf{x}$-data points) by multiplication with $(\mathbf{I} - \mathbf{W}\mathbf{P}^t)$ from the right, it is sufficient to do a single filtering of the $\mathbf{X}_o$ regression coefficients $\boldsymbol{\beta}_o = a\mathbf{w}_1$ by multiplication with $(\mathbf{I} - \mathbf{W}\mathbf{P}^t)$ from the left to obtain the regression coefficients $\boldsymbol{\beta}$ to be applied for the original unfiltered data.

The regression coefficients $\boldsymbol{\beta}$ in this case are clearly a linear combination including both $\mathbf{w}_1$ and the $\mathbf{W}$'s in (9) that are associated with $\mathbf{y}$-orthogonality. In contrast to the regression coefficients $\mathbf{b}$ in the NAP/DO modelling approach, the $\boldsymbol{\beta}$ of Fearns method is not "blind" w.r.t. the alleged irrelevant phenomena in $\mathbf{W}$ that are associated with $\mathbf{y}$-orthogonality.

2.2 The O-PLS of Trygg and Wold

Fearn both formulated and solved an optimization problem to justify his method (including a solution to the 'new sample problem' issued in [1, section 5.2]). Nevertheless, Fearns solution to the OSC-problem was soon overtaken by a heuristic approach, i.e. the patented O-PLS algorithm of Trygg and Wold [2] that was introduced with the following explanation:

"...The proposed O-PLS method analyzes the disturbing variation in each regular PLS component. The non-correlated variation in $\mathbf{X}$ is separated from the correlated variation, with the additional benefit that the non-correlated variation itself can be studied and analyzed. Removing non-correlated variation in data prior to data modeling is not only interesting from a predictive point of view, but the interpretational ability of resulting models also improves. Thus more information and knowledge of a system can be retrieved and analyzed, and developed further."

O-PLS (with its offsprings proposed in [14] and [15]) soon became, and still is the most popular choice for OSC modeling and calibration, see Pinto et al. [16].

The O-PLS algorithm presented in [2, section 2.3] is a recipe for filtering (preprocessing) the $\mathbf{X}$-data using slightly different $\mathbf{y}$-orthogonal scores and associated loading weights than those found by Fearns approach. The original formulation of O-PLS does not provide a transparent procedure for calculating a corresponding regression model (only the calculations for the $\mathbf{y}$-orthogonal filtering factors and associated weights is described). The steps for calculating regression coefficients are instead left for a possible subsequent application of ordinary PLS to the filtered $\mathbf{X}$-data.

From later insights, in particular given by Ergon [4], Indahl [17] and the equations (2) and (3) above, the most important characteristics of a k-component O-PLS model (including the calculation of regression coefficients) are the following:

An O-PLS model with $k \geq 2$ components has sets of orthogonal weights $\{\mathbf{v}_1, ..., \mathbf{v}_k\}$ and associated orthogonal scores $\{\mathbf{t}_1, ..., \mathbf{t}_k\}$ where

- The weights $\mathbf{v}_1 = -\mathbf{w}_2, \ ... \ , \mathbf{v}_{k-1} = -\mathbf{w}_k, \mathbf{v}_k = \mathbf{w}_1$, where $\mathbf{w}_1, \ ... \ , \mathbf{w}_k$ are the ordinary PLS-weights (according to Ergon [4]).

- The first $(k-1)$ score vectors $\{\mathbf{t}_1, ... \ , \mathbf{t}_{k-1}\}$ are $\mathbf{y}$-orthogonal, i.e. $\mathbf{t}_i^t \mathbf{y} = 0$ for $1 \leq i \leq k-1$, and they span the same subspace as the non-orthogonal PLS-scores $\{\mathbf{X}\mathbf{w}_2, ... \ , \mathbf{X}\mathbf{w}_k\}$ (that are $\mathbf{y}$-orthogonal according to equation (3)).

- The $\mathbf{y}$-orthogonal filtering of $\mathbf{X}$ is

$$\mathbf{X}_o = \mathbf{X} - \mathbf{T}\mathbf{P}^t = (\mathbf{I} - \mathbf{T}(\mathbf{T}^t\mathbf{T})^{-1}\mathbf{T}^t)\mathbf{X} = \mathbf{X}(\mathbf{I} - \mathbf{V}(\mathbf{P}^t\mathbf{V})^{-1}\mathbf{P}^t). \tag{15}$$

  In matrix notation, the $\mathbf{y}$-orthogonal scores $\mathbf{T} = [\mathbf{t}_1 \ ... \ \mathbf{t}_{k-1}]$ satisfy the identity

$$\mathbf{T} = \mathbf{X}\mathbf{V}(\mathbf{P}^t\mathbf{V})^{-1}, \tag{16}$$

  where

$$\mathbf{V} = [\mathbf{v}_1 \ ... \ \mathbf{v}_{k-1}] = -[\mathbf{w}_2 \ ... \ \mathbf{w}_k], \tag{17}$$

  and $\mathbf{P}^t = (\mathbf{T}^t\mathbf{T})^{-1}\mathbf{T}^t\mathbf{X}$ (according to Indahl [17, section 3]).

- The last (k-th) score vector $\mathbf{t}_k = \mathbf{X}_o\mathbf{v}_k$ is obtained by a Gram-Schmidt step filtering off the $(k-1)$ $\mathbf{y}$-orthogonal factors in $\mathbf{X}\mathbf{v}_k$, i.e.

$$\mathbf{t}_k = (\mathbf{I} - \mathbf{T}(\mathbf{T}^t\mathbf{T})^{-1}\mathbf{T}^t)\mathbf{X}\mathbf{v}_k = \mathbf{X}_o\mathbf{v}_k, \tag{18}$$

  where $\mathbf{I}$ is the identity matrix.

- The complete set of scores $\{\mathbf{t}_1, \ ... \ , \mathbf{t}_k\}$ coincide with the set of orthogonal vectors obtained by an application of the Gram-Schmidt orthogonalization process (QR-factorization) with the following sequence of non-orthogonal score vectors: $\mathbf{X}\mathbf{v}_1, \ ..., \ \mathbf{X}\mathbf{v}_k$ (according to Indahl [17, section 3 and appendix A.1]).

- The fitted values of an O-PLS model with $(k-1)$ $\mathbf{y}$-orthogonal components is $\hat{\mathbf{y}} = \alpha\mathbf{t}_k$, where α is the least squares solution of $a\mathbf{t}_k = \mathbf{y}$ and from (18) the $\mathbf{X}_o$-regression coefficients are

$$\boldsymbol{\beta}_o = \alpha\mathbf{v}_k \ (= \alpha\mathbf{w}_1). \tag{19}$$

- From (15), (16) and (19) it follows that the fitted values can alternatively be expressed as

$$\hat{\mathbf{y}} = \mathbf{X}_o(\alpha\mathbf{v}_k) = \mathbf{X}(\mathbf{I} - \mathbf{V}(\mathbf{P}^t\mathbf{V})^{-1}\mathbf{P}^t)\boldsymbol{\beta}_o = \mathbf{X}\boldsymbol{\beta}, \tag{20}$$

  where the $\mathbf{X}$-regression coefficients

$$\boldsymbol{\beta} = (\mathbf{I} - \mathbf{V}(\mathbf{P}^t\mathbf{V})^{-1}\mathbf{P}^t)\boldsymbol{\beta}_o \tag{21}$$

coincide with the regression coefficients obtained by a k-component application of ordinary PLS to the $(\mathbf{X}, \mathbf{y})$-data.

- The columns of $\mathbf{T}$ are orthogonal to $\mathbf{X}_o$ (by construction). Therefore the $\mathbf{y}$-orthogonal scores $\{\mathbf{t}_1, \ldots, \mathbf{t}_{k-1}\}$ are also $\hat{\mathbf{y}}$-orthogonal, i.e.

$$\mathbf{T}^t\hat{\mathbf{y}} = \mathbf{T}^t\mathbf{X}_o\boldsymbol{\beta}_o = (\mathbf{T}^t\mathbf{X} - \mathbf{T}^t\mathbf{T}\mathbf{P}^t)\boldsymbol{\beta}_o$$

$$= (\mathbf{T}^t\mathbf{X} - \mathbf{T}^t\mathbf{T}(\mathbf{T}^t\mathbf{T})^{-1}\mathbf{T}^t\mathbf{X})\boldsymbol{\beta}_o = (\mathbf{T}^t\mathbf{X} - \mathbf{T}^t\mathbf{X})\boldsymbol{\beta}_o = \mathbf{0}. \tag{22}$$

Because the eliminated $\mathbf{T}$-part resulting in $\mathbf{X}_o$ is $\mathbf{y}$-orthogonal, the identity $\mathbf{X}_o^t\mathbf{y} = \mathbf{X}^t\mathbf{y} = c\mathbf{w}_1$ holds. PLS applied to the $(\mathbf{X}_o, \mathbf{y})$-data will therefore produce $\mathbf{w}_1$ as the first vector of loading weights, and then subsequent $\mathbf{w}$'s orthogonal to $\mathbf{w}_1$. By equation (2) these subsequent $\mathbf{w}$'s result in $\mathbf{y}$-orthogonal scores $\mathbf{tw}$ that should have been filtered off $\mathbf{X}$ (or equivalently off $\mathbf{X}\mathbf{v}_k$) in the first place. Consequently, the recommendation in [2] for applying ordinary PLS to the filtered data $\mathbf{X}_o$ is redundant.

Just like equation (14) in Fearns approach, we also see that rather than filtering the present $\mathbf{X}$-matrix or new $\mathbf{x}$-data points by multiplication with the matrix $(\mathbf{I} - \mathbf{V}(\mathbf{P}^t\mathbf{V})^{-1}\mathbf{P}^t)$ from the right, it is sufficient to do just the single filtering of $\boldsymbol{\beta}_o$ in (21) by multiplication with this particular matrix from the left.

The model spaces spanned by the non-orthogonal scores $\{\mathbf{X}\mathbf{v}_1, \ldots, \mathbf{X}\mathbf{v}_k\}$ and $\{\mathbf{X}\mathbf{w}_1 \ldots \mathbf{X}\mathbf{w}_k\}$ must necessarily coincide. Because the first set of vectors span the model space of O-PLS and the second set span the model space of the ordinary PLS, the introduction of the O-PLS algorithm in [2] is clearly superfluous from a model fitting point of view. The only difference between PLS and O-PLS is that their common model space is represented by two alternative choices of orthogonal score basis vectors $\mathbf{T}$ and associated $\mathbf{P}$-loadings. Navigation between such alternative bases is always a simple task. More on the technical details of the equivalence between ordinary PLS and O-PLS is given in [17, section 3].

Finally we note that the $\mathbf{X}$-regression coefficient vector $\boldsymbol{\beta}$ of O-PLS (21) is a linear combination of the ordinary PLS loading weights $\mathbf{w}_1, \ldots, \mathbf{w}_k$ where all except $\mathbf{w}_1$ are associated with $\mathbf{y}$-orthogonality. Unlike the regression coefficients $\mathbf{b}$ from the NAP/DO modelling approach, the $\boldsymbol{\beta}$ of O-PLS is not "blind" w.r.t. the alleged irrelevant phenomena to be associated with $\mathbf{y}$-orthogonality. Above, we saw that the same thing was true for Fearns method.

2.3 Ordinary PLS and y-orthogonality

In the Chemometrics community (as confirmed by the above quote from Wold et al. [1, section 5]), one usually describes ordinary PLS as a method maximizing of the covariance between $\mathbf{X}$ and $\mathbf{y}$. However, except for the first component, this is true only in the deflated sense of $\mathbf{X}$ and $\mathbf{y}$. In terms of the undeflated $\mathbf{X}$, most PLS algorithms explicitly calculate the orthogonal loading weights $\mathbf{w}_i$ and corresponding non-orthogonal scores $\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1$ and $\mathbf{t}_i^\star = \mathbf{X}\mathbf{w}_i$ ($2 \leq i \leq k$, that are also $\mathbf{y}$-orthogonal) prior to a Gram-Schmidt step to assure mutual orthogonality of the scores.

Although O-PLS and PLS generate exactly the same structure from a dataset in terms of subspaces, identical orthogonal loading weights (when ignoring differences in signs as indicated above) [4], $\mathbf{X}$-regression coefficients and model fits, the resulting "models" are equipped with very different heuristics (proposed by the inventors of these methods) regarding both information content and other interpretations.

In spite of the re-arrangement of the non-orthogonal PLS-scores prior to computing the orthogonal scores of O-PLS, there is really no rigorous justification for two such interpretation alternatives. In this context it should be noted that with k components, there are $k!$ possible ways of permuting the non-orthogonal scores $\mathbf{Xw}_1, ..., \mathbf{Xw}_k$, leading to $k!$ different orthogonal bases for exactly the same subspace. If $\mathbf{T}_\pi = [\mathbf{t}_{\pi(1)} \; ... \; \mathbf{t}_{\pi(k)}]$ is the orthogonal basis obtained by applying the Gram-Schmidt procedure to the permuted (by some permutation π of the numbers $1, ..., k$) non-orthogonal scores $\mathbf{Xw}_{\pi(1)}, ..., \mathbf{Xw}_{\pi(k)}$, the associated matrix of loadings is $\mathbf{P}_\pi = \mathbf{X}^t\mathbf{T}_\pi(\mathbf{T}_\pi\mathbf{T}_\pi)^{-1}$. This means that for each loading weight $\mathbf{w}_i$ and associated (non-orthogonal) score $\mathbf{Xw}_i$ there are a large number of alternative associated scores $\mathbf{t}_{\pi(i)}$ (and corresponding loadings $\mathbf{p}_{\pi(i)}$ subject to the widely accepted PLS/O-PLS interpretation heuristics) related to some orthogonal basis.

However, the order of <u>deriving</u> the loading weights $\mathbf{w}_1, ..., \mathbf{w}_k$ inside the respective algorithms is identical for both the PLS and the O-PLS. According to Wold et al. [18], these weight vectors correspond to the conjugate (orthogonal) gradient directions for generating the solution of the normal equations

$$\mathbf{X}^t\mathbf{Xb} = \mathbf{X}^t\mathbf{y} \tag{23}$$

associated with the OLS problem. By equation (2) their mutual orthogonality is equivalent to the $\mathbf{y}$-orthogonality of the non-orthogonal scores $\mathbf{Xw}_i$ for $i \geq 2$.

It should be noted that the presentation of the NIPALS PLS in [18] also included a deflation step for $\mathbf{y}$. Björck [19] has criticized the omitted $\mathbf{y}$-deflation in the more recent applications of the NIPALS PLS, because this introduces an unnecessary and possibly harmful loss of numerical precision in the resulting PLS-solutions. If the more recent algorithms derived from the NIPALS PLS had not ignored the numerically favourable $\mathbf{y}$-deflation, one could rightfully wonder if inventions such as the O-PLS algorithm would have been made at all.

Additional (non-orthogonal) components $\mathbf{Xw}_i$ contributing to reducing the residual $\mathbf{y}$, improves the fit of the original $\mathbf{y}$ accordingly. The same vectors (or some particular orthogonal basis derived from them) can not alternatively be taken as an explanation of what has nothing to do with $\mathbf{y}$ without introducing a contradiction.

Regarding $\mathbf{y}$-orthogonality of the non-orthogonal PLS scores, the sceptical reader is strongly encouraged to compute the $\mathbf{t}_i^\star = \mathbf{Xw}_i$ for $i \geq 2$ (either from the NIPALS algorithm or directly by the non-orthogonal and $\mathbf{y}$-deflating PLS algorithm of Martens) for <u>any</u> dataset to verify empirically their $\mathbf{y}$-orthogonality.

2.4 PLS+ST and the Target Projection

2.4.1 The PLS+ST

The *PLS post-processing by similarity transformation* (PLS+ST) proposed by Ergon [4] represent an alternative way of computing the desired O-PLS score vector $\mathbf{t}_k$ given in equation (18).

Let $\mathbf{t}_1 = \mathbf{Xw}_1 (= \mathbf{Xv}_k)$ and let $\mathbf{T}^\star = [\mathbf{Xw}_2 \; ... \; \mathbf{Xw}_k]$ be the matrix representation of the non-orthogonal (and $\mathbf{y}$-orthogonal) PLS-scores. To obtain an even simpler expression for the O-PLS score vector $\mathbf{t}_k$, Ergon took advantage of the $\mathbf{y}$-orthogonal part $\mathbf{T}^\star\mathbf{q}_2$ of the expression for the fitted values in Martens non-orthogonal scores PLS. The resulting simplified expression for the fitted values is given by

$$\hat{\mathbf{y}} = \mathbf{XWq} = q_1\mathbf{t}_1 + \mathbf{T}^\star\mathbf{q}_2, \tag{24}$$

where $\mathbf{W} = [\mathbf{w}_1 \ ... \ \mathbf{w}_k]$ is the matrix representation of the PLS loading weights. The regression coefficient vector $\mathbf{q} = [q_1 \ \mathbf{q}_2^t]^t$ is associated with the score vectors in (24), and the sub-vector $\mathbf{q}_2 = [q_2 \ ... \ q_k]^t$ represent the regression coefficients associated with the $(k-1)$ $\mathbf{y}$-orthogonal score vectors. Ergon noted that the desired O-PLS score vector $\mathbf{t}_k$ in (18) also can be expressed as

$$\mathbf{t}_k = q_1^{-1}\hat{\mathbf{y}} = \mathbf{t}_1 + q_1^{-1}\mathbf{T}^\star\mathbf{q}_2, \tag{25}$$

where the last term in (25) represent the required $\mathbf{y}$-orthogonal correction of $\mathbf{t}_1$.

From the definitions $\mathbf{w}_{PLS+ST} = \mathbf{w}_1 + q_1^{-1}\sum_{i=2}^{k} q_i\mathbf{w}_i$ and $\mathbf{W}_2 = [\mathbf{w}_2 \ ... \ \mathbf{w}_k]$, the PLS+ST transformed loading weights and associated scores are given by $\mathbf{W}_\mathbf{M} = \mathbf{WM} = [\mathbf{w}_{PLS+ST} \ \mathbf{W}_2]$ and $\mathbf{T}_\mathbf{M} = \mathbf{XW}_\mathbf{M} = [\mathbf{t}_k \ \mathbf{T}^\star]$, respectively, where $\mathbf{M} = \begin{bmatrix} 1 & \mathbf{0} \\ q_1^{-1}\mathbf{q}_2 & \mathbf{I} \end{bmatrix}$ is the required transformation matrix with the inverse $\mathbf{M}^{-1} = \begin{bmatrix} 1 & \mathbf{0} \\ -q_1^{-1}\mathbf{q}_2 & \mathbf{I} \end{bmatrix}$. By a simple manipulation of (24) we have

$$\hat{\mathbf{y}} = \mathbf{XW}(\mathbf{MM}^{-1})\mathbf{q} = \mathbf{T}_\mathbf{M}\mathbf{M}^{-1}\mathbf{q} = \mathbf{T}_\mathbf{M}\mathbf{q}_\mathbf{M} = q_1\mathbf{t}_k = q_1\mathbf{Xw}_{PLS+ST}, \tag{26}$$

because the $\mathbf{T}_\mathbf{M}$-regression coefficients $\mathbf{q}_\mathbf{M} = (\mathbf{M}^{-1}\mathbf{q}) = [q_1 \ 0 \ ... \ 0]^t$. The associated $\mathbf{X}$-regression coefficients

$$\boldsymbol{\beta} = \mathbf{Wq} = q_1\mathbf{w}_{PLS+ST} \tag{27}$$

required for predictions of new $\mathbf{x}$-data points, must necessarily coincide with the expression in equation (21).

It should be noted that the (first) vector $\mathbf{w}_{PLS+ST}$ of transformed loading weights is neither a unit vector nor orthogonal to the other columns in $\mathbf{W}_\mathbf{M}$, and that only the associated transformed score vector ($\mathbf{t}_k = \mathbf{Xw}_{PLS+ST}$) is orthogonal to the other $\mathbf{T}_\mathbf{M}$-columns (the $\mathbf{y}$-orthogonal scores $\mathbf{T}^\star$).

Due to the non-orthogonalities in $\mathbf{W}_\mathbf{M}$, the proposed notation in [4, equation 6] for the rank k approximation $\mathbf{X}_k$ of $\mathbf{X}$ is quite confusing, i.e. $\mathbf{X}_k = \mathbf{t}_k\mathbf{w}_1^t + \mathbf{T}^\star(\mathbf{W}_2 - q_1^{-1}\mathbf{w}_1\mathbf{q}_2^t)^t$, and

$$\mathbf{X} = \mathbf{t}_k\mathbf{w}_1^t + \mathbf{T}^\star(\mathbf{W}_2 - q_1^{-1}\mathbf{w}_1\mathbf{q}_2^t)^t + \mathbf{E} = \mathbf{X}_k + \mathbf{E},$$

where $\mathbf{E}$ represent the residual part. Note that a much simpler expression for the approximation part $\mathbf{X}_k$ is available, i.e.

$$\mathbf{X}_k = \mathbf{T}_\mathbf{M}\mathbf{M}^{-1}\mathbf{W}^t = \mathbf{XWMM}^{-1}\mathbf{W}^t = \mathbf{XWW}^t.$$

A compact view of the PLS+ST can be obtained directly by considering equation (27) for an ordinary PLS model:

1. Define the desired vector of loading weights $\mathbf{w}_{PLS+ST} = q_1^{-1}\boldsymbol{\beta}$

2. The vector of fitted values

$$\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} = q_1\mathbf{X}\mathbf{w}_{PLS+ST} = q_1\mathbf{t}_k = \mathbf{H}\mathbf{y}, \tag{28}$$

   where $\mathbf{H}$ is the (symmetric) orthogonal projection onto the PLS model space (which is spanned by the non-orthogonal scores $\mathbf{XW} = [\mathbf{t}_1\ \mathbf{T}^\star]$).

3. Because the $\mathbf{y}$-orthogonal columns of $\mathbf{T}^\star$ form a subset of the $\mathbf{XW}$-columns, the projection matrix $\mathbf{H}$ must satisfy $\mathbf{HT}^\star = \mathbf{T}^\star$ and

$$\hat{\mathbf{y}}^t\mathbf{T}^\star = (\mathbf{Hy})^t\mathbf{T}^\star = \mathbf{y}^t(\mathbf{HT}^\star) = \mathbf{y}^t\mathbf{T}^\star = \mathbf{0}. \tag{29}$$

Equation (29) shows that the $\mathbf{y}$-orthogonal column vectors in $\mathbf{T}^\star$ are also $\hat{\mathbf{y}}$- (and $\mathbf{t}_k$-) orthogonal. The latter simply means that $\mathbf{y}$-orthogonal filtering of the fitted values $\hat{\mathbf{y}}$ (obtained by PLS) with respect to $\mathbf{T}^\star$ to improve the prediction ability of a model, is just as sensible as sending a healthy patient to ineffective surgery.

2.4.2 The Target Projection (TP) method

The description of the PLS+ST method presented in [4, section 3] has no particular focus on model interpretations. This is, however, included in the equivalent TP method. According to Kvalheim [5], the vector of TP-loading weights is defined as the unit vector

$$\mathbf{w}_{TP} = \|\boldsymbol{\beta}\|^{-1}\boldsymbol{\beta}, \tag{30}$$

and from (27) we have $\mathbf{w}_{TP} = (q_1\|\boldsymbol{\beta}\|^{-1})\mathbf{w}_{PLS+ST}$. The corresponding TP-score vector is given by

$$\begin{aligned}\mathbf{t}_{TP} &= \mathbf{X}\mathbf{w}_{TP} \\ &= \|\boldsymbol{\beta}\|^{-1}\mathbf{X}\boldsymbol{\beta} = \|\boldsymbol{\beta}\|^{-1}\hat{\mathbf{y}},\end{aligned} \tag{31}$$

and the associated vector of loadings often preferred for interpretations is

$$\begin{aligned}\mathbf{p}_{TP} &= \mathbf{X}^t\mathbf{t}_{TP}/(\mathbf{t}_{TP}^t\mathbf{t}_{TP}) = \mathbf{X}^t\mathbf{t}_{TP}/\|\mathbf{t}_{TP}\|^2 \\ &= (\|\mathbf{t}_{TP}\|^2\|\boldsymbol{\beta}\|)^{-1}\mathbf{X}^t\hat{\mathbf{y}}.\end{aligned} \tag{32}$$

In the full rank OLS case (obtained by including the maximum number of PLS components), we have $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}$, where the regression coefficient vector $\boldsymbol{\beta}$ is found by solving the associated normal equations

$$\begin{gathered}\mathbf{X}^t\mathbf{X}\mathbf{b} = \mathbf{X}^t\mathbf{y}(= c\mathbf{w}_1) \\ \Downarrow \\ \mathbf{X}^t\hat{\mathbf{y}} = \mathbf{X}^t\mathbf{y}.\end{gathered} \tag{33}$$

with respect to $\mathbf{b}$. Thus for OLS, the normal equations (33) implies that the TP-loading vector $\mathbf{p}_{TP}$ in (32) is proportional to the familiar (first) PLS loading weight vector $\mathbf{w}_1$:

$$\mathbf{p}_{TP} = (\|\mathbf{t}_{TP}\|^2\|\boldsymbol{\beta}\|)^{-1}\mathbf{X}^t\hat{\mathbf{y}} = (\|\mathbf{t}_{TP}\|^2\|\boldsymbol{\beta}\|)^{-1}\mathbf{X}^t\mathbf{y} = c(\|\mathbf{t}_{TP}\|^2\|\boldsymbol{\beta}\|)^{-1}\mathbf{w}_1, \tag{34}$$

where $c = \|\mathbf{X}^t\mathbf{y}\|$ is the normalization constant associated with $\mathbf{w}_1$. Kvalheim [5, equation 10] recognized this relationship by an alternative route without noticing the underlying normal equations (33), and instead concluded:

"*Thus, in the absence of truncation, the TP loadings are proportional to the PLS weights on the first PLS component. This is an important result since* $\mathbf{w}_1$ *represents the normalized co-variance vector between* $\mathbf{X}$ *and* $\mathbf{y}$ *(in variable space) ...*

...The PLS weights on the first component, $\mathbf{w}_1$ *might be a better choice for revealing the* $\mathbf{x}$*-variables most influential with respect to the response since the TP loadings, representing the co-variances between the* $\mathbf{x}$*-variables and the predicted response, converge towards these weights (Equation (10)). However, as we shall see when we look further into this matter ... this choice is still not optimal.*

... one may conclude that the TP loadings are most appropriate for revealing the predictive part of $\mathbf{X}$*. Alternatively, one may use the PLS weights* $\mathbf{w}_1$ *since the TP loadings converge towards these weights when the variation in* $\mathbf{X}$ *is exhausted. However, none of these vectors represent an optimal choice.*"[5].

Note that neither the OLS-regression coefficient vector leading to the perfect fit of the right hand side $c\mathbf{w}_1$ in the normal equations (33), nor the PLS-regression coefficients in (21) approximately solving (33), are considered to be appropriate candidates for revealing the predictive part of $\mathbf{X}(= [\mathbf{x}_1 \ ... \ \mathbf{x}_p])$.

In [5], the more "optimal choice" is claimed to be obtained by the so-called *selectivity ratios* (SR). Before stating the SR-definition, one should note that by introducing the diagonal scaling matrix

$$\mathbf{S}_c = \|\mathbf{t}_{TP}\| \begin{bmatrix} \|\mathbf{x}_1\|^{-1} & ... & 0 \\ \vdots & \ddots & \vdots \\ 0 & ... & \|\mathbf{x}_p\|^{-1} \end{bmatrix}, \tag{35}$$

the vector of so-called *correlation loadings*, $\mathbf{r}_{TP}$ corresponding to the vector $\mathbf{p}_{TP}$ of TP loadings, is given by

$$\mathbf{r}_{TP} = \mathbf{S}_c\mathbf{p}_{TP}. \tag{36}$$

The correlation between the i-th variable $\mathbf{x}_i$ and $\hat{\mathbf{y}}$, is of course identical to the correlation between $\mathbf{x}_i$ and $\mathbf{t}_{TP}$, and its value is $r_i = \mathbf{r}_{TP}(i)$, i.e. the i-th entry of the correlation loading vector $\mathbf{r}_{TP}$.

The *selectivity ratio* SR_i of the i-th variable $\mathbf{x}_i$ is defined in [5, equation 11] as the ratio between the explained and the unexplained variances when regressing the i-th variable $\mathbf{x}_i$ onto $\mathbf{t}_{TP}$. In terms of the correlation r_i, this definition is equivalent to

$$SR_i = SR(r_i^2) = r_i^2/(1 - r_i^2), \tag{37}$$

In [5] the SR definition is introduced with the following explanation:

"*... Thus, the individual SR_is are closely related to the correlation between the predictive part of an* $\mathbf{x}$*-variable and the response* $\mathbf{y}$*, but division with the unexplained variance produces a more sensitive measure than the correlation. Furthermore, the SR provides a bridge from the co-variance-based TP loadings to a variance-independent measure without the deteriorative effect of noise from small variables accompanying the scaling of the* $\mathbf{x}$*-variables to unit variance. ...*"

Note that the first sentence before the comma in the above quote is consistent with the mathematical definition of the SR_i only for OLS models. For PLS models, the correct relationship is to the correlation between the $\mathbf{x}$-variables and the fitted values $\hat{\mathbf{y}}$ (not $\mathbf{y}$). The text after the comma is misleading because the function

$$SR(t) = t/(1-t) \tag{38}$$

from the definition (37) is strictly increasing for arguments $t \in [0, 1)$, i.e. there is a one-to-one correspondence in the relationship between the squared correlations and the selectivity ratios. The second sentence in the quote is misleading by the same one-to-one correspondence, where no moderation of spurious correlations (due to small $\mathbf{x}$-variances) takes place.

The precise relationship between the TP loadings $\mathbf{p}_{TP}$ in (32) and the SR-values, is given by the correlation loadings $\mathbf{r}_{TP}$ in (36) and the function $SR(\cdot)$ defined in (38). When approaching full rank in the PLS-modelling (when getting close to the OLS-model) the entries of the vector $\mathbf{r}_{TP}$ will become close to the univariate correlations between $\mathbf{y}$ and the $\mathbf{x}$-variables that can be calculated directly from the data (prior to the regression modelling).

2.5 A second look at Fearns OSC alternative

The $\mathbf{y}$-orthogonal matrix $\tilde{\mathbf{X}}$ obtained by the deflation in equation (4) is maximal in terms of rank, and its rank is only one less than the rank of $\mathbf{X}$. Therefore one might wrongly conclude that this matrix represents the maximum amount of information (in terms of rank) not needed to model $\mathbf{y}$. However,

$$\mathbf{X}_1 = \mathbf{X} - \tilde{\mathbf{X}} = \mathbf{X}\mathbf{w}_1\mathbf{w}_1^t \tag{39}$$

is a (filtered) matrix of rank 1 that accounts for the residual information in $\mathbf{X}$ not being orthogonal to $\mathbf{y}$. Regressing $\mathbf{y}$ onto the rank 1 matrix $\mathbf{X}_1$ results in a vector of fitted values that are proportional to $\mathbf{X}_1\mathbf{w}_1 = \mathbf{X}\mathbf{w}_1 = \mathbf{t}_1$, and deflation of $\mathbf{X}_1$ with respect to $\mathbf{t}_1$ results in a $\mathbf{0}$-matrix. Hence, we are captured in the situation of an ordinary PLS model based on one component only. By directly eliminating the $\mathbf{W}$-directions (the entire row space of $\tilde{\mathbf{X}}$) associated with $\mathbf{y}$-orthogonality, we throw away the information necessary to improve on the residual $\mathbf{y}$'s. Both Fearns method and the O-PLS identifies and use such $\mathbf{W}$-directions (obtained by different strategies though) in subsequent order to obtain models comparable in both approach and performance to PCR and PLS, respectively. Without using this possibility of correcting the initial $\mathbf{t}_1$ for its major $\mathbf{y}$-orthogonal components, that are associated with directions of significant variance in the column space of $\tilde{\mathbf{X}}$, neither of the two methods would work beyond the first component ($\mathbf{t}_1$).

Because there is no general way of initially "guessing" a good score vector without significant

$\mathbf{y}$-orthogonal components, a poorer "guess" (some linear combination of the $\mathbf{X}$-columns - usually $\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1$) must be taken as the starting point. Reducing the $\mathbf{y}$-residuals either

- by directly introducing components $(\mathbf{X}\mathbf{w}_2, ..., \mathbf{X}\mathbf{w}_k)$ subsequent to $\mathbf{t}_1$ for obtaining better projections of $\mathbf{y}$, or

- by eliminating exactly the same ($\mathbf{y}$-orthogonal) components from the $\mathbf{y}$-orthogonal component of $\mathbf{t}_1$

is just two sides of the same coin. In the deflated sense of $\mathbf{y}$, i.e. by continuing the modelling process beyond the first component $\mathbf{t}_1$, the associated $\mathbf{y}$-residual $(\mathbf{r}_{i-1})$ and the corresponding component $\mathbf{X}\mathbf{w}_i$ $(2 \leq i \leq k)$ are of course not orthogonal, and an improved model fit is therefore obtained.

As explained above, both Fearns OSC and the O-PLS are consistent with this strategy. Fearn applies a selection of the dominant and $\mathbf{y}$-orthogonal left singular vectors of $\tilde{\mathbf{X}}$ into a PCR flavoured solution. The O-PLS use the $\mathbf{y}$-orthogonal PLS-scores. For both methods it must be stressed that the subsequently derived $\mathbf{y}$-orthogonal directions are not orthogonal to the corresponding $\mathbf{y}$-residuals (the $\mathbf{X}\mathbf{w}_i$'s are non-orthogonal to the $\mathbf{r}_{i-1}$'s for $1 \leq i \leq k$). This explains precisely how including more $\mathbf{y}$-orthogonal directions in the model building is working when the $\mathbf{y}$-deflations are ignored.

2.5.1 An informative computer experiment

In the field of chemometrics, there is a long tradition in modifying various established algorithms (in particular the NIPALS PLS) as part of the research process towards new data analysis methods. Several of the published OSC-modelling approaches have obviously evolved in this way.

According to this well established tradition, the readers are therefore encouraged to do some relevant computer experiments. In particular you should try to verify that the fitting of a response vector $\mathbf{y}$ can be approached by orthogonalizing an arbitrary initial guess with respect to $\mathbf{y}$-orthogonal directions as follows:

- Pick your favourite NIR- (or any other) dataset $\mathbf{X}$ with n rows, p columns and an associated response vector $\mathbf{y} \in \mathbb{R}^n$ (mean centring of $\mathbf{y}$ and the $\mathbf{X}$-columns should be included).

- Generate a random vector $\mathbf{w} \in \mathbb{R}^p$ and compute the corresponding *random linear combination* of the $\mathbf{X}$-columns, i.e. $\mathbf{t} = \mathbf{X}\mathbf{w}$, as your initial $\hat{\mathbf{y}}$-guess. Then adjust this $\mathbf{t}$ according to:

    1. Fearns strategy (improve $\mathbf{t}$ by subtracting its projection onto the first 10 left singular vectors of the $\mathbf{y}$-orthogonal matrix $\tilde{\mathbf{X}}$ defined in (4) - use equation (11) where you replace $\mathbf{t}_1$ with your random guess $\mathbf{t}$).
    2. PLS (improve $\mathbf{t}$ by subtracting its projection onto the subspace spanned by the first 10 $\mathbf{y}$-orthogonal scores, i.e. $\mathbf{t}_2^\star = \mathbf{X}\mathbf{w}_2$,...,$\mathbf{t}_{11}^\star = \mathbf{X}\mathbf{w}_{11}$).

- Compute the correlation between $\mathbf{y}$ and the improved $\mathbf{t}$ in both cases.

- Repeat 1. and 2. starting with $\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1$ as your initial $\hat{\mathbf{y}}$-guess, and compare the correlations.

In all the cases above you should be able to observe that the proposed correlations are highly similar.

3 Examples with y-orthogonality and imposed orthogonality constraints in the sample space

3.1 A published case where the O-PLS idea actually fails to work

According to the introduction in [2], the OSC-issues being solved by the O-PLS are

"...remove systematic information in $\mathbf{X}$ *not correlated to the modelling of* $\mathbf{y}$ *in order to achieve better models in multivariate calibration. ...Its objective is to improve interpretations of PLS models and reduce model complexity. O-PLS provides a way to remove ... variability in* $\mathbf{X}$ *that is orthogonal to* $\mathbf{y}$*. The proposed O-PLS method analyzes the disturbing variation in each regular PLS component. The non-correlated variation in* $\mathbf{X}$ *is separated from the correlated variation, with the additional benefit that the non-correlated variation itself can be studied and analyzed. Removing non-correlated variation in data prior to data modeling is not only interesting from a predictive point of view, but the interpretation ability of resulting models also improves. ..."*

A simulated example given by Trygg and Wold [2, section 2.3.10] is intended to illustrate some of the O-PLS capabilities, and it deserves a careful investigation:

In this example we consider two closely related data matrices $\mathbf{X}_0$, $\mathbf{X}_1$ (to obtain exact precision for those who like hand calculations, two extra digits have been included in each entry of $\mathbf{X}_1$) and one response vector $\mathbf{y}$:

$$\mathbf{X}_0 = \begin{bmatrix} -1 & -1 \\ 1 & -1 \\ -1 & 1 \\ 1 & 1 \end{bmatrix}, \ \mathbf{X}_1 = \begin{bmatrix} -2.1825 & -2.1825 \\ 1.8375 & -0.1625 \\ -0.4825 & 1.5175 \\ 0.8275 & 0.8275 \end{bmatrix} \ \text{and} \ \mathbf{y} = \begin{bmatrix} 2 \\ 2 \\ 0 \\ -4 \end{bmatrix}$$

Note that $\mathbf{X}_1$ is a corrupted version of $\mathbf{X}_0$ obtained by adding the $\mathbf{y}$-orthogonal vector $\mathbf{t}_{orth} = [-1.1825,\ 0.8375,\ 0.5175,\ -0.1725]^t$ to the $\mathbf{X}_0$-columns, and that the identity $\mathbf{X}_0^t\mathbf{y} = \mathbf{X}_1^t\mathbf{y} = [-4, -8]^t$ holds. According to the quoted introduction of [2], one might be tempted to expect the O-PLS to eliminate the effects of the disturbances caused by the vector $\mathbf{t}_{orth}$. On the other hand, we know that ordinary least squares (OLS), PLS with two components (full rank), O-PLS and Fearns method (both of full rank due to extraction and filtering of the data with respect to one $\mathbf{y}$-orthogonal component) necessarily must result in identical models (i.e. identical regression coefficients).

For $(\mathbf{X}_0, \mathbf{y})$, the vector of least squares regression coefficients (found by all the four methods) is $\mathbf{b}_0 = [-1, -2]^t$. For $(\mathbf{X}_1, \mathbf{y})$, the four methods also agree and the vector of least squares regression coefficients is $\mathbf{b}_1 = [0.083, -1.0758]^t$. The corresponding residual vectors

$$\mathbf{r}_0 = \mathbf{X}_0\mathbf{b}_0 - \mathbf{y} \text{ and } \mathbf{r}_1 = \mathbf{X}_1\mathbf{b}_1 - \mathbf{y}$$

have norms $\|\mathbf{r}_0\| = 2.0000$ and $\|\mathbf{r}_1\| = 3.9656$, respectively. Consequently, the $\mathbf{y}$-orthogonality introduced by adding $\mathbf{t}_{orth}$ to the original $\mathbf{X}_0$-columns has lead to a poorer fit in the second model obtained by both O-PLS, Fearns method and (of course) OLS.

In [2] however, it is wrongly claimed that O-PLS is able to remove the effect of $\mathbf{y}$-orthogonal (non-correlated) variation in $\mathbf{X}_1$! The problem here is that the introduced $\mathbf{y}$-orthogonal vector $\mathbf{t}_{orth}$ is neither contained in the column space of $\mathbf{X}_0$ nor in the column space of the corrupted matrix $\mathbf{X}_1$, i.e. there is no $\mathbf{w} \in \mathbb{R}^2$ so that the identity $\mathbf{t}_{orth} = \mathbf{X}_1\mathbf{w}$ holds. Only external information about the vector $\mathbf{t}_{orth}$ is really helpful in eliminating its influence to obtain a better model fit in this particular case.

If external knowledge of $\mathbf{t}_{orth}$ were available, there would still be some subtle issues to clarify. Subtraction of $\mathbf{t}_{orth}$ from the $\mathbf{X}_1$-columns is obviously not the same thing as a filtering of the $\mathbf{X}_1$-matrix to make its columns orthogonal to $\mathbf{t}_{orth}$. Furthermore, in this particular example it is easily shown that the vector $\mathbf{t}_{orth}$ is non-orthogonal even to the original $\mathbf{X}_0$-columns. Therefore, orthogonalizing the $\mathbf{X}_1$-columns with respect to $\mathbf{t}_{orth}$ would not bring back the "uncontaminated" $\mathbf{X}_0$. Instead we would obtain the following matrix:

$$\mathbf{X}_2 = (\mathbf{I} - \mathbf{t}_{orth}(\mathbf{t}_{orth}^t\mathbf{t}_{orth})^{-1}\mathbf{t}_{orth}^t)\mathbf{X}_1 = \begin{bmatrix} -0.3440 & -0.6596 \\ 0.5354 & -1.2411 \\ -1.2871 & 0.8510 \\ 1.0957 & 1.0497 \end{bmatrix}. \tag{40}$$

Note that $\mathbf{X}_2$ will also be the result of filtering $\mathbf{X}_0$ in the same way (just replace $\mathbf{X}_1$ by $\mathbf{X}_0$ in equation (40)), and that the first ordinary vectors of PLS loading weights for all these matrices are identical, i.e. $\mathbf{X}_2^t\mathbf{y}(= \mathbf{X}_0^t\mathbf{y} = \mathbf{X}_1^t\mathbf{y}) = [-4, -8]^t$.

Solving the $(\mathbf{X}_2, \mathbf{y})$ regression problem by OLS or any of the methods PLS, O-PLS or Fearns OSC (full rank versions) would result in the least squares regression coefficients $\mathbf{b}_2 = [-1.4908, -2.2546]^t$. The corresponding residual norm in this case would be $\|\mathbf{X}_2\mathbf{b}_2 - \mathbf{y}\| = 0$, i.e. a perfect fit.

The perfect fit (hardly intended when the example was prepared for [2]) is a mathematical consequence of working with centered data: The two linearly independent $\mathbf{X}_2$-columns are orthogonal to both the constant vector $\mathbf{1} = [1\ 1\ 1\ 1]^t$ and $\mathbf{t}_{orth}$ (so is also $\mathbf{y}$). Due to their linear independence, the $\mathbf{X}_2$ columns together with the vectors $\mathbf{1}$ and $\mathbf{t}_{orth}$ represent a basis for $\mathbb{R}^4$. This means that any vector in $\mathbb{R}^4$ (including $\mathbf{y}$) can be represented as a linear combination of these four basis vectors. Because $\mathbf{y}$ is orthogonal to both $\mathbf{1}$ and $\mathbf{t}_{orth}$, both its coefficients (coordinates) associated with these vectors must be 0. Consequently, $\mathbf{y}$ is perfectly represented by a linear combination of the two $\mathbf{X}_2$-columns!

3.2 Model interpretations in least squares modelling is a challenging task

Improved model interpretation is one of the most important "selling" points for the O-PLS methodology. Brown and Green [20], however, stressed that even for the relatively simple class of least squares regression methods, model interpretation may be a challenging subject. It is indeed much more challenging than what may be the impression from traditional chemometrics using PLS and O-PLS.

It is in fact not too hard to derive orthogonal models that both fit (and predict) a dataset well. In the example below we demonstrate this fact using a MATLAB benchmark dataset (available from MATLAB's *Statistics and Machine Learning Toolbox* [21] by the command: `'load spectra'`) of NIR/octane measurements. A complete description of the dataset is given in [22].

We start by finding the solution $\hat{\boldsymbol{\beta}}$ minimizing the constrained least squares problem

$$\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|^2 + \lambda\|\mathbf{L}_1\boldsymbol{\beta}\|^2, \tag{41}$$

with $\lambda = 0.1$ and $\mathbf{L}_1$ denoting the discrete 1. derivative operator. The second term in (41) penalizes roughness in the solution vector $\hat{\boldsymbol{\beta}}$. Note that in this case, $\hat{\boldsymbol{\beta}}$ is a linear combination of rows in the augmented matrix $\begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{L}_1 \end{bmatrix}$.

With the solution $\hat{\boldsymbol{\beta}}$ at hand, we may seek an alternative PLS-based model where the regression coefficient $\hat{\mathbf{b}}$ is constrained to be orthogonal to the solution $\hat{\boldsymbol{\beta}}$ of (41) (this can be done by orthogonalizing the rows of $\mathbf{X}$ with respect to $\hat{\boldsymbol{\beta}}$ prior to the PLS modelling). The resulting orthogonal vector of regression coefficients and scatter plots of the corresponding leave-one-out cross validation (CV) predictions $\hat{\mathbf{y}}_{cv}$ are shown in Figure 2.

Note that neither $\hat{\mathbf{b}}$ nor $\hat{\boldsymbol{\beta}}$ are forced to be exact linear combinations of the original $\mathbf{X}$-rows, but both alternatives do work as intended for the original $\mathbf{X}$-measurements. Although the two models are almost indistinguishable from a predictive point of view (see the bottom part of Figure 2), it seems quite impossible to provide a trustworthy interpretation simultaneously accounting for the two orthogonal regression coefficient vectors $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{b}}$ (see the upper part of Figure 2).

4 Discussion

The example in section 3.1 shows very clearly that strange things can happen if the $\mathbf{X}$-columns are manipulated by some arbitrary $\mathbf{y}$-orthogonal vector $\mathbf{t}_{orth} \in \mathbb{R}^n$ not contained in the column space of $\mathbf{X}$. Even with exact knowledge of $\mathbf{t}_{orth}$, our clues about the handling of new data points $\mathbf{x} \in \mathbb{R}^p$ for later predictions would be very limited. Unfortunately, extensions of the O-PLS methodology such as the O2-PLS [14] and the OnPLS [15] cannot save us from this peculiar situation.

O-PLS, PLS+ST and TP have in common that they identify the subspace spanned by the ordinary PLS scores, and that they introduce a new set of basis vectors for this subspace. In all these alternatives, the (first) basis vector of real interest is chosen in the direction of the fitted values $\hat{\mathbf{y}}$ (the other basis vectors are chosen to be orthogonal to both $\hat{\mathbf{y}}$ and $\mathbf{y}$). In this

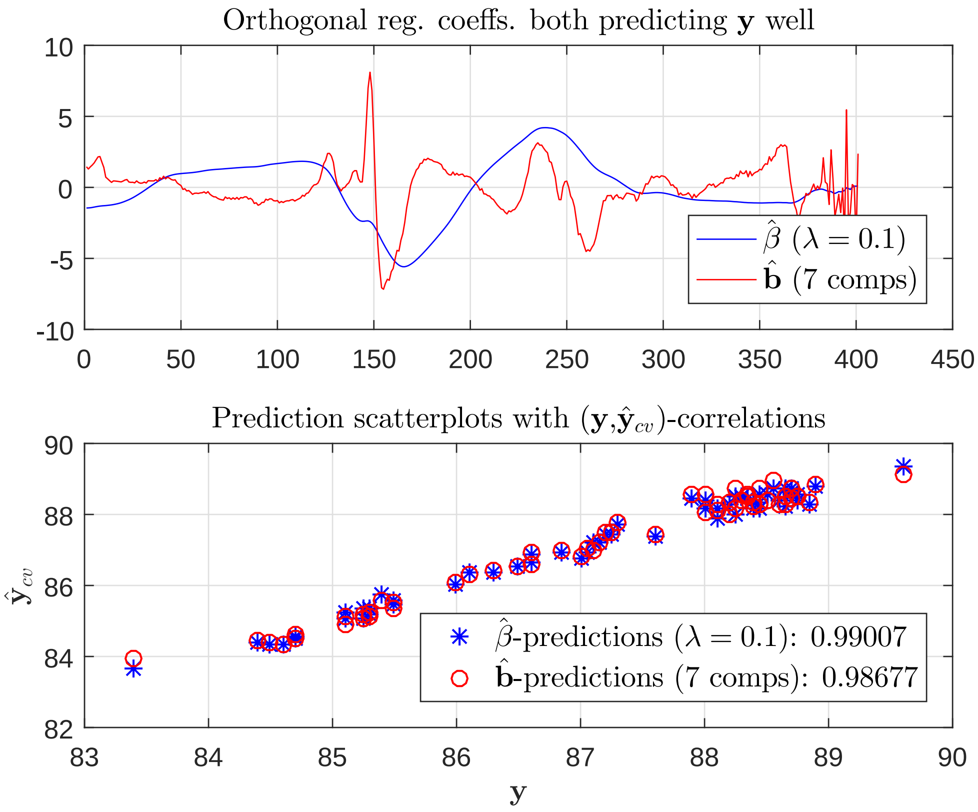

Figure 2: Orthogonal regression coefficients from Tikhonov-regularization ($\mathbf{L}_1$ and $\lambda = 0.1$) and constrained PLS (7 components) giving similar predictions.

perspective, these equivalent methods seems to be superfluous constructions made to emphasize (and interpret) $\mathbf{y}$-orthogonality and/or to justify some simplified model interpretation heuristics based on the fitted values $\hat{\mathbf{y}}$ (in some scaled version) as the score vector of main interest.

In the user community of PLS-methodology, inspection of the so-called $\mathbf{p}$-loadings resulting from the NIPALS algorithm (and the O-PLS algorithm proposed in [2]) is considered as a vitally important part of the model interpretation heuristics. The $\mathbf{p}$-loadings relate to a particular orthogonal basis of the column subspace. As discussed in section 2.3, there are a large number of possibilities ($k!$ - including the two bases obtained by PLS and O-PLS) for choosing such bases for the subspace spanned by the non-orthogonal PLS-scores. Each such possibility leads to a particular matrix of associated $\mathbf{p}$-loadings with a corresponding set of possible interpretations. Linking an underlying non-orthogonal score vector $\mathbf{X}\mathbf{w}_i$ to the orthogonal (in some basis) score vector $\mathbf{t}_{\pi(i)}$ and the interpretations based on the associated $\mathbf{p}_{\pi(i)}$-loading is clearly a risky and ambiguous business, because of the large number of different possibilities for obtaining the various possible $\mathbf{t}_{\pi(i)}$-scores.

The precise mathematical meaning of a loading vector $\mathbf{p}$ is most eaily seen by considering the orthogonal projection of the $\mathbf{X}$-columns onto the subspace spanned by the score vector $\mathbf{t}$ (of some orthognal basis):

$$\hat{\mathbf{X}} = \mathbf{t}(\mathbf{t}^t\mathbf{t})^{-1}\mathbf{t}^t\mathbf{X} = \mathbf{t}\mathbf{p}^t, \tag{42}$$

where $\mathbf{p}^t = \mathbf{t}^t\mathbf{X}/\mathbf{t}^t\mathbf{t}$ by definition is the (transposed) vector of $\mathbf{p}$-loadings associated with $\mathbf{t}$. From (42) the i-th entry p_i of $\mathbf{p}$ clearly corresponds to the $\mathbf{t}$-coordinate of the projected i-th column of $\mathbf{X}$. If $\mathbf{t}$ is chosen as a unit vector (i.e. $\mathbf{t}^t\mathbf{t} = 1$, a choice corresponding to using an orthonormal basis), we realize that the vector $\mathbf{p}(= \mathbf{X}^t\mathbf{t})$ also corresponds to the direction maximizing the $(\mathbf{X}, \mathbf{t})$-covariance.

As a comment to the example in section 3.1, where the "arbitrary" $\mathbf{y}$-orthogonal vector $\mathbf{t}_{orth}$ was introduced, a simple projection argument shows that any of the p column vectors $\mathbf{x}_i$ of $\mathbf{X}$ can be projected onto $\mathbf{y}$ and expressed as

$$\mathbf{x}_i = c_i\mathbf{y} + \mathbf{r}_i, \text{ where } \mathbf{r}_i^t\mathbf{y} = 0 \text{ (i.e. } \mathbf{y}\text{-orthogonal) and } c_i \in \mathbb{R}.$$

If some clever OSC algorithm where capable of eliminating all these $\mathbf{y}$-orthogonal $\mathbf{r}_i$'s (that are not necessarily linear combinations of the $\mathbf{X}$-columns), the resulting filtered data matrix would look like

$$\mathbf{X}_{oo} = [c_1\mathbf{y}\ c_2\mathbf{y}\ ...c_p\mathbf{y}] = \mathbf{y}\mathbf{c}^t$$

where the only information in $\mathbf{X}$ not completely lost are the coefficients $\mathbf{c}^t = [c_1\ ...c_p] = (\mathbf{y}^t\mathbf{y})^{-1}\mathbf{y}^t\mathbf{X}$ proportional to the covariances between $\mathbf{y}$ and the $\mathbf{X}$-columns, i.e.

$$\mathbf{X}^t\mathbf{y} = \mathbf{X}_{oo}^t\mathbf{y}.$$

In the introduction we emphasized that the present focus on OSC methods is restricted to the single response case. In the multi-response case with $\mathbf{Y} \in \mathbb{R}^{n\times q}$ $(q \geq 2)$, an $\mathbf{Y}$-orthogonal vector is still defined as a linear combination $\mathbf{t} = \mathbf{X}\mathbf{w}$ of the $\mathbf{X}$-columns satisfying

$$\mathbf{t}^t\mathbf{Y} = \mathbf{w}^t(\mathbf{X}^t\mathbf{Y}) = \mathbf{0}, \tag{43}$$

i.e. the vector $\mathbf{w} \in \mathbb{R}^p$ of coefficients is orthogonal to the subspace spanned by the columns of $\mathbf{X}^t\mathbf{Y}$. Let's assume that the matrix $\mathbf{W}_1 \in \mathbb{R}^{p\times q}$ represent an orthonormal basis for this subspace ($\mathbf{W}_1$ can be obtained either by QR-factorization, or the "thin" SVD, of $\mathbf{X}^t\mathbf{Y}$). Then equation (43) is equivalent to requiring $\mathbf{w}^t\mathbf{W}_1 = \mathbf{0}$. Deflation of $\mathbf{X}$ with respect to $\mathbf{W}_1$ results in the $\mathbf{Y}$-orthogonal matrix

$$\tilde{\mathbf{X}} = \mathbf{X} - (\mathbf{X}\mathbf{W}_1)\mathbf{W}_1^t. \tag{44}$$

Note that $\underline{\text{any}}$ algorithm (including the O2-PLS in [14]) finding $\underline{\text{any}}$ $\mathbf{Y}$-orthogonal vector $\mathbf{t} = \mathbf{X}\mathbf{w}$, automatically finds a linear combinations of the $\tilde{\mathbf{X}}$-columns because $\mathbf{W}_1^t\mathbf{w} = \mathbf{0}$ by the remarks after (43), and

$$\mathbf{t} = \mathbf{X}\mathbf{w} = \mathbf{X}\mathbf{w} - (\mathbf{X}\mathbf{W}_1)\mathbf{W}_1^t\mathbf{w} = \tilde{\mathbf{X}}\mathbf{w} \tag{45}$$

by (44). In particular, the multi-response case described in Fearn [13], and the associated maximization problem, is solved by finding the dominant right singular vector of $\tilde{\mathbf{X}}$ in (44). Like in the single response case discussed in section 2.1, the additional factors are given by the subsequent right singular vectors ordered by the associated singular values.

For any linear regression method (single- or multiblock) modelling the q responses of $\mathbf{Y}$ by a k-dimensional space V (assuming $k > q$), the fitted values $\hat{\mathbf{Y}}$ will account for q dimensions in V, and the $\mathbf{Y}$-orthogonal vectors will account for the remaining $(k-q)$ dimensions. By considering Figure 1 (ignore its sub-text) for this purpose, one may think of the q dimensions of $\hat{\mathbf{Y}}$ as collapsed into the "line" spanned by $\hat{\mathbf{Y}}$. The $\mathbf{z}$ then represents a vector in the $(k-q)$-dimensional (rather than $(k-1)$) subspace of $\mathbf{Y}$-orthogonal vectors. Thus, for any single- or multiblock linear regression method with one or several responses, the concept of $\mathbf{Y}$-orthogonality is always available.

There is an important difference between the O-PLS and the NAP/DO approach in considering the interferents (the vectors in $\mathbb{R}^p$ representing the phenomena explained as irrelevant to the particular $\mathbf{y}$ of interest). In O-PLS the filtered matrix $\mathbf{X}_o$ in (15) is obtained by multiplication of $\mathbf{X}$ from the right hand side with the skew (oblique) projection matrix $(\mathbf{I}-\mathbf{V}(\mathbf{P}^t\mathbf{V})^{-1}\mathbf{P}^t)$, where the matrix $\mathbf{V}$ represent the irrelevant phenomena. Note that the rows of $\mathbf{X}_o$ are indeed orthogonal to $\mathbf{V}$ because $(\mathbf{I}-\mathbf{V}(\mathbf{P}^t\mathbf{V})^{-1}\mathbf{P}^t)\mathbf{V}=\mathbf{0}$. The problem is, however, that the $\mathbf{X}_o$-regression coefficients $\boldsymbol{\beta}_o$ in (19) are sensitive to the filtering operation, i.e. $(\mathbf{I}-\mathbf{V}(\mathbf{P}^t\mathbf{V})^{-1}\mathbf{P}^t)\boldsymbol{\beta}_o = \boldsymbol{\beta}(\neq \boldsymbol{\beta}_o)$, where $\boldsymbol{\beta}$ is the ordinary PLS regression coefficients according to equation (21). The latter means that the entire filtering of $\mathbf{X}$ into $\mathbf{X}_o$ is collapsed into $\boldsymbol{\beta}$. Therefore, these regression coefficients (and not $\boldsymbol{\beta}_o$) must be applied to the original unfiltered data $\mathbf{X}$. Although $\boldsymbol{\beta}_o$ is "blind" to the $\mathbf{V}$-directions in $\mathbf{X}$, it is unfortunately not orthogonal to the $\mathbf{P}$-loadings, i.e. the skew projection $\mathbf{V}(\mathbf{P}^t\mathbf{V})^{-1}\mathbf{P}^t\boldsymbol{\beta}_o \neq \mathbf{0}$. The regression coefficient vector $\boldsymbol{\beta}$ on the other hand, is a linear combination of both $\mathbf{w}_1$ and the $\mathbf{V}$-vectors) and takes the $\mathbf{V}$-directions in the $\mathbf{X}$-rows into account for predicting the analyte/phenomenon (represented by $\mathbf{y}$) well. In the end, this means that considering (or even interpreting) the $\mathbf{V}$-vectors as real interferents must be incorrect. This is not the case with the regression coefficients $\mathbf{b}$ obtained from the NAP/DO approach described in section 2. Equation (6) demonstrates that the proposed matrix $\mathbf{P}_a$ of interferents and corresponding directions in the unfiltered $\mathbf{X}$-rows do not interact with the regression coefficients $\mathbf{b}$.

Before a successful elimination of the effects of some (a) interferents/phenomena that are potentially present in the sample signals (such as NIR-spectra), some kind of prior knowledge on how to establish the particular $\mathbf{P}_a$-matrix is required. Thereafter, elimination of the effects associated with $\mathbf{P}_a$ can either be obtained by "blinding" the samples (rows in $\mathbf{X}$) with respect to $\mathbf{P}_a$ before modelling (as described in the NAP/DO approach above), or by constraining the regression coefficient vector $\mathbf{b}$ to be orthogonal to the interferents/phenomena in $\mathbf{P}_a$ as an integrated part of the modelling.

Modelling with integration of such constraints can be managed well by the Tikhonov regularization (TR) approach to linear regression model building. In the TR approach, elimination of influence by the (a) disturbing interferents collected in $\mathbf{P}_a$ is handled as follows:

- Define the associated extended data matrix, response vector and regression problem as

$$\mathbf{X}_e = \begin{bmatrix} \mathbf{X} \\ \sqrt{\mu}\mathbf{P}_a^t \\ \sqrt{\lambda}\mathbf{I} \end{bmatrix}, \mathbf{y}_e = \begin{bmatrix} \mathbf{y} \\ \mathbf{0}_a \\ \mathbf{0}_p \end{bmatrix} \text{ and } \mathbf{X}_e\mathbf{b} = \begin{bmatrix} \mathbf{X}\mathbf{b} \\ \sqrt{\mu}\mathbf{P}_a^t\mathbf{b} \\ \sqrt{\lambda}\mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{y} \\ \mathbf{0}_a \\ \mathbf{0}_p \end{bmatrix} = \mathbf{y}_e,$$

  respectively. Here $\mathbf{I}$ is the $p \times p$ identity matrix, $\sqrt{\mu}$ and $\sqrt{\lambda}$ are positive regularization constants to be chosen according to the purpose of the modelling, and $\mathbf{b}$ is the unknown solution of the augmented regression problem. Note that the extended data matrix $\mathbf{X}_e$ is guaranteed to have full rank due to its bottom block $\sqrt{\lambda}\mathbf{I}$.

- The least squares expression to be minimized w.r.t. $\mathbf{b}$ is the residual norm

$$\|\mathbf{X}_e\mathbf{b}-\mathbf{y}_e\|^2 = \|\mathbf{X}\mathbf{b}-\mathbf{y}\|^2 + \mu\|\mathbf{P}_a^t\mathbf{b}-\mathbf{0}_a\|^2 + \lambda\|\mathbf{I}\mathbf{b}-\mathbf{0}_p\|^2 = \|\mathbf{X}\mathbf{b}-\mathbf{y}\|^2 + \mu\|\mathbf{P}_a^t\mathbf{b}\|^2 + \lambda\|\mathbf{b}\|^2,$$

  where the parameters μ and λ are assumed to be fixed.

- By choosing μ to be (very) large, we see that the solution $\hat{\mathbf{b}}$ of the above least squares problem can be forced to be as close to orthogonal to the columns of $\mathbf{P}_a$ as we like. An appropriate choice of λ has the effect of preventing the squared norm $\|\hat{\mathbf{b}}\|^2$ of the regression coefficients from blowing up without an unnecessary sacrifice of precision in the fitting $\mathbf{X}\hat{\mathbf{b}}$ of $\mathbf{y}$. The latter is well known as the bias-variance trade off problem in ridge regression (RR).

- The normal equations and the associated unique OLS solution $\hat{\mathbf{b}}$ of the augmented system are

$$\mathbf{X}_e^t\mathbf{X}_e\mathbf{b} = \mathbf{X}_e^t\mathbf{y}_e \Leftrightarrow (\mathbf{X}^t\mathbf{X} + \mu\mathbf{P}_a^t\mathbf{P}_a + \lambda\mathbf{I})\mathbf{b} = \mathbf{X}^t\mathbf{y} \Rightarrow \hat{\mathbf{b}} = (\mathbf{X}^t\mathbf{X} + \mu\mathbf{P}_a^t\mathbf{P}_a + \lambda\mathbf{I})^{-1}\mathbf{X}^t\mathbf{y},$$

  respectively. Note that no "blinding" or direct filtering of the $\mathbf{X}$-data takes place in TR.

The survey on the TR methodology given by Kalivas [23], is recommended as a more detailed explanation on the necessities and possibilities available with this type of approach.

5 Conclusions (by points)

Is O-PLS correcting or confusing? The question was asked in the title of this paper, and answers are provided by the following summary points:

1. The idea of OSC as proposed in [1] and the most popular algorithm for doing OSC (the O-PLS proposed in [2]) were, according to the authors, to develop a methodology for improving predictions and interpretations. The fact that ordinary PLS already works according to this principle, i.e. by identifying subsequent components orthogonal to $\mathbf{y}$ for improving the model fit, or equivalently reducing the residual $\mathbf{y}$, was ignored by the authors of [1] and [2].

2. Single response models obtained by PLS and O-PLS are actually in one-to-one correspondence because the algorithms are equivalent and find exactly the same subspace for fitting the response $\mathbf{y}$.

   Conclusions of 1. and 2: Incorrect and confusing - there is no improvement in model fit and predictions compared to PLS. The O-PLS also introduced alternative interpretation heuristics, essentially building on the PLS-loading weights $\mathbf{w}_2$, ..., $\mathbf{w}_k$, that were in conflict with the traditional interpretation heuristics of PLS (building on the same loading weights). By overlooking that the O-PLS is equivalent to the ordinary PLS, the O-PLS inventors contributed to more confusion by introducing the alternative model interpretation heuristics.

3. An example in [2, section 2.3.9] being designed to demonstrate the capabilities of O-PLS, fails in filtering out the $\mathbf{y}$-orthogonal information that was deliberately added to the $\mathbf{X}$-data. The example, with calculations done correctly (shown in section 3.1 above), actually demonstrates the limitations of the entire OSC-idea.

Conclusion: The desired orthogonal signal correction fails in removing the effects of $\mathbf{y}$-orthogonal factors from outside the column space of the desired $\mathbf{X}$-data.

4. Figure 1 in the introduction shows that $\mathbf{y}$-orthogonality is always present as a part of multiple linear regression modelling. For the orthogonal projection $\hat{\mathbf{Y}}$ of $\mathbf{Y} \in \mathbf{R}^{n \times q}$ onto some k-dimensional $(q < k \leq n)$ model space $V \subseteq \mathbf{R}^n$, there are always $(k - q)$ dimensions in V that are orthogonal to both $\hat{\mathbf{Y}}$ and $\mathbf{Y}$.

   Conclusion: By realizing that this always is the case (in particular for OLS, PCR and PLS with $q = 1$), much confusion regarding the meaning of $\mathbf{Y}$-orthogonal filtering in OSC could have been avoided.

5. In the discussion part comparing the O-PLS and the NAP/DO approach, we were forced to conclude that O-PLS cannot help us to find any real interferents. This is so because the only possible choice of regression coefficients applicable for the unfiltered data $\mathbf{X}$, is the ordinary PLS regression coefficients $\boldsymbol{\beta}$. This means that sensitivity of the $\mathbf{V}$-directions defined in (17) (and the associated non-orthogonal scores $\mathbf{X}\mathbf{w}_2, ..., \mathbf{X}\mathbf{w}_k$) is required from the $\mathbf{X}$-regression coefficients to obtain a good fit of of $\mathbf{y}$.

   In section 2.3 the entire concept of $\mathbf{y}$-orthogonality, as proposed in OSC and identified by O-PLS, is explained by its unfortunate ignorance of the always present $\mathbf{y}$-residuals, and their relevance for explaining the phenomenon/analyte associated with $\mathbf{y}$. The fact that subsequent $\mathbf{y}$-residuals are always non-orthogonal to the corresponding sequence of non-orthogonal scores is ignored by OSC/O-PLS. The $\mathbf{V}$-vectors defining these scores, are precisely the conjugate gradient directions required for subsequently reducing the $\mathbf{y}$-residuals in the search for a good model explaining the phenomenon/analyte represented by $\mathbf{y}$. The missing deflation of $\mathbf{y}$ in the more recent implementations of the NIPALS PLS algorithm is probably the main reason why unfortunate confusions regarding $\mathbf{y}$-orthogonality in OSC has become so wide spread.

   Conclusion: Speaking of $\mathbf{y}$-orthogonality and $\mathbf{y}$-orthogonal components without requiring orthogonality in terms of the associated $\mathbf{y}$-residuals is not good model interpretation practice. Successful modelling including the context of "irrelevant" interferents/phenomena requires some prior knowledge about how their associated directions can be identified and represented. Both NAP/DO and the described TR approach represent sound model building strategies resulting in models that are "blinded" in the directions of the provided irrelevant phenomena/interferents. Therefore, later filtering of new datapoints are not required. The models obtained by the O-PLS (and Fearns method) do not have this property - their $\mathbf{X}$-regression coefficients are forced to be non-orthogonal to the alleged irrelevant directions.

6 Acknowledgements

The present paper is an invited contribution to this spesial issue in Jounal of Chemometrics based on the author's earlier contributions in [17] and an oral presentation given at the SSC14-meeting (Sardinia, 2015).

I am grateful for a considerable number of critical remarks given by an anonymous referee. These remarks contributed significantly to improving the level of precision in the manuscript, and to sharpen its conclusions.

The Research Council of Norway (project number 239070) provided financial support for this work.

References

[1] Wold S, Antti H, Lindgren F, Öhman J. Orthogonal signal correction of near-infrared spectra. Chemom. Intell. Lab. Syst. 1998; 44: 175-185.

[2] Trygg J, Wold S. Orthogonal projections to latent structures, O-PLS. J. Chemom. 2002; 16: 119-128.

[3] Anderson M. A comparison of nine PLS1 algorithms. J. Chemom. 2009; 23: 518-529.

[4] Ergon R. PLS post-processing by similarity transformation (PLS+ST): a simple alternative to OPLS. J. Chemom. 2005; 96: 1-4.

[5] Kvalheim, OM. Interpretation of partial least squares regression models by means of target projection and selectivity ratio plots. J. Chemom. 2010; 24: 496-504.

[6] Kemsley EK, Tapp HS. OPLS filtered data can be obtained directly from non-orthogonalized PLS1. J. Chemom. 2009; 23(5-6): 263-264.

[7] Sjöblom J, Svensson O, Josefson M, Kullberg H, Wold S. An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra. Chemom. Intell. Lab. Syst. 1998; 44: 229-244.

[8] Andersson CA. Direct orthogonalization. Chemom. Intell. Lab. Syst. 1999; 47: 51-63.

[9] Westerhuis J, de Jong S, Smilde AK. Direct orthogonal signal correction Chemom. Intell. Lab. Syst. 2001; 56: 13-25.

[10] Svensson O, Kourti T, McGregor JF. An investigation of orthogonal signal correction algorithms and their characteristics. J. Chemom. 2002; 16: 176-188.

[11] Goicoechea HC, Olivieri AC. A comparison of orthogonal signal correction and net analyte preprocessing methods. Theoretical and experimental study. Chemom. Intell. Lab. Syst. 2001; 56: 73-78.

[12] Ni W, Brown SD, Man R. The relationship between net analyte signal/preprocessing and orthogonal signal correction algorithms. Chemom. Intell. Lab. Syst. 2009; 98: 97-107.

[13] Fearn, T. On orthogonal signal correction. Chemom. Intell. Lab. Syst. 2000; 50: 47-52.

[14] Trygg J, Wold S. O2-PLS, a two-block (X-Y) latent variable regression (LVR)methodwith an integral OSC filter. J. Chemom. 2003; 17: 53-64.

[15] Löfstedt T, Trygg J. OnPLS-a novel multiblock method for the modelling of predictive and orthogonal variation. J. Chemom. 2011; 25: 441-455.

[16] Pinto RC, Trygg J, Gottfries J. Advantages of orthogonal inspection in chemometrics. J. Chemom. 2012; 26: 231-235.

[17] Indahl UG. Towards a complete identification of orthogonal variation in multiple regression from a PLS1 modeling point of view: including OPLS by a change of orthogonal basis. J. Chemom. 2014; 28: 508-517.

[18] Wold S, Ruhe A, Wold H, Dunn WJ. The collinearity problem in linear regression, the partial least squares (PLS) approach to generalized inverses. SIAM J. Sci. Stat. Comput. 1984; 5: 735-743.

[19] Björck, Å. Stability of two direct methods for bidiagonalization and partial least squares. SIAM J. Matrix Anal. Appl. 2014; 35(1): 279-291.

[20] Brown CD and Green RL. Critical factors limiting the interpretation of regression vectors in multivariate calibration. Trends in Analytical Chemistry, 2009; 28(4): 506-514.

[21] MATLAB R2016a. Statistics and Machine Learning Toolbox.
www.mathworks.com/products/statistics/

[22] Kalivas JH. Two Data Sets of Near Infrared Spectra. Chemom. Intell. Lab. Syst. 1997; 37: 255-259.

[23] Kalivas JH. Overview of two-norm (L2) and one-norm (L1) Tikhonov regularization variants for full wavelength or sparse spectral multivariate calibration models or maintenance. J. Chemom. 2012; 26: 218-230.